REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Oct. 13, 1997	Quarterly Tech Report

4. TITLE AND SUBTITLE
Scientific And Technical Report
Intelligent Metacomputing Testbed

5. FUNDING NUMBERS
F19 628-96-C-0020

6. AUTHOR(S)
Jon Genetti Reagan Moore
Richard Marciano

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
San Diego Supercomputer Center
P0 Box 85608
San Diego CA 92186-5608

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ARPA/ITO
Att: Gary Koob
3701 N. Fairfax Drive
Arlington VA 22203-1714

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

19971017 121

12a. DISTRIBUTION/AVAILABILITY STATEMENT
ARPA Agent
ARPA/ITO
DCAA San Diego North County Office

DTIC QUALITY INSPECTED 2

13. ABSTRACT (Maximum 200 words)

The Distributed Object Computation Testbed (DOCT) has two principal goals: the demonstration of an object computation environment that supports distributed processing of large archived data sets and the demonstration of support for electronic submission and processing of complex documents and patent applications for the U.S. Patent and Trademark Office (USPTO). The infrastructure that is being integrated to create this testbed includes archival storage systems, databases, an object computation system, document management systems, and intelligent agents that support the patent application workflow. The resulting technologies should also apply to the information needs of other agencies, such as the National Science Foundation, the National Institutes of Health, the Nuclear Regulatory Commission, the Environmental Protection Agency (EPA), the Department of Energy, and the Department of Defense.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Data Handling, Data Mining, Distributed Computing	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

Intelligent Metacomputing Testbed
(Distributed Object Computational Testbed (DOCT))

San Diego Supercomputer Center

Reagan Moore, Principal Investigator

QUARTERLY SCIENTIFIC & TECHNICAL REPORT

July 1997 - September 1997

Sponsored by:

Advanced Research Projects Agency/ITO

Arpa Order No. D570

Issued by ESC/ENS under contract F19628-96-c-0020

Table of Contents

1. Task Objectives

The Distributed Object Computation Testbed (DOCT) has two principal goals: the demonstration of an object computation environment that supports distributed processing of large archived data sets, and the demonstration of support for electronic submission and processing of complex documents and patent applications for the U.S. Patent and Trademark Office (USPTO). The infrastructure that is being integrated to create this testbed includes archival storage systems, databases, an object computation system, document management systems, and intelligent agents that support the patent application workflow. The resulting technologies should also apply to the information needs of other agencies, such as the National Science Foundation, the National Institutes of Health, the Nuclear Regulatory Commission, the Environmental Protection Agency (EPA), the Department of Energy, and the Department of Defense.

The DOCT project consists of a collaboration of eight research organizations, led by the San Diego Supercomputer Center. The participating organizations include:

- California Institute of Technology (Caltech)
- National Center for Supercomputing Applications (NCSA)
- Old Dominion University (ODU)
- Open Text Corporation (Open Text)
- Science Applications International Corporation (SAIC)
- University of California at San Diego (UCSD)
- University of Virginia (UVa)

The DOCT project is investigating the creation of an innovative, distributed, national-scale, persistent document handling that will be capable of manipulating, searching, and managing terabytes of compound, complex, mixed-mode documents and data sets. This system provides the underlying infrastructure for supporting the document management services needed for electronic commerce applications. The DOCT hardware/network systems consists of high performance computing, communication, and storage devices distributed nationwide at university and government sites from coast to coast.

The DOCT software infrastructure is being created by integration of the following software layers and associated functions:

- Intelligent agents.
- Document management system.
- Applications.
- Persistent object computation system.
- Application-level scheduling system.
- Communication system.
- Data handling system.
- Object-relational database management system.
- Archival storage system.

The tasks listed in this report are defined in a Research and Development Plan and Schedule, accessible at the URL:

http://www.sdsc.edu/DOCT/Publications/rdps-doc.html

2. Technical Problems

The major network performance problem has been improving achievable bandwidth between SDSC, SAIC and Caltech. The round-trip message time is about 70 milliseconds from SDSC to SAIC across the AAInet/ATDnet network. Maximizing the bandwidth utilization requires keeping enough messages in flight such that message receipt acknowledgments do not cause delays in the data flow. Upgrades have been made to the Solaris operating system on the CS6400 at SAIC to support a larger number of packets in flight. This has resulted in sustained measured bandwidths of 48 Mbps. Routers along the path begin queuing packets at higher transmission rates. A report on this effort is available at the DOCT URL.

The second network problem is connectivity. The vBNS connection to Caltech is still not in place, although it is approved. This is required to support replication of data between two archives.

When the networks are fully functional, we will be moving 200 GB of data between the sites for the replicated archive and for the distributed analysis of document data type conformance.

3. General Methodology

3.1 Technical Methodology

The software infrastructure is being assembled through the integration of metacomputing environments (Legion) with distributed data handling technology and electronic commerce systems. The latter systems are being built on top of commercial software packages, including workflow environments and commercial databases.

- The intelligent agent framework has been designed using the TexCel workflow and document management system.
- The text search system is based on Opentext.
- Applications are being built on database technology to support tasks such as claims graph dependency.
- The persistent object computation system is being formed from Legion, with version 0.5 released in July 1997. Version 1.0 is planned for November, 1997.
- The data handling system is extending the technology developed in the DARPA funded MDAS project (http://www.sdsc.edu/MDAS). In particular, a Storage Resource Broker (SRB) is being used to integrate Legion access to archived storage. A release of the SRB software was provided in September, 1997.
- A distributed authentication/encryption systems has been developed to minimize risk from spoofing of servers within the DOCT testbed. The system also provides mechanisms for self-registration of users of the system.
- The archival storage system has been implemented using the production High Performance Storage System (HPSS) at both SDSC and Caltech.

The integration of these systems is shown in Figure 4-1.

3.2 Distributed Environment

The distributed environment that comprises the DOCT testbed is shown in Figure 3-1. This shows each of the participating sites and the network connections that link the sites together.

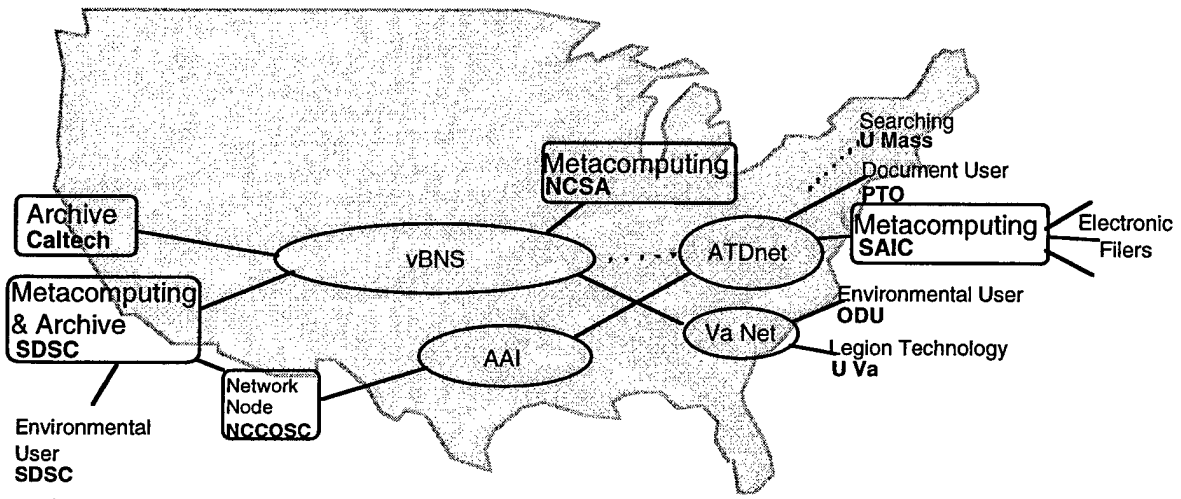

Figure 3-1: Distributed Object Computation Testbed network interconnects

3.3 Management Methodology

Management is through weekly teleconferences, planned workshops and quarterly demos.

Teleconferences that have been held are listed below, with the corresponding task identifiers from the Research and Development Plan and Schedule:

- July 7: C2-4,C3-1 (HPSS replication), C5-1,C5-5 (MDAS status)
- July 14: D3,D4 (Wrappers for Livelink search, Texcel)
- July 28: A14 (Classification), A5-2 (Validation)
- August 4: DOCT network-based security
- August 11: F1-2,F3-3 (Queuing systems), E5-2 (Fault Tolerance)
- August 18: G2-2/3, G3-1/2 (Agent Framework Status), C5-1/2/3/4/5 (MDAS Status)
- August 25: A10-3 (Environmental Demo), D1-2/3/4 (Vault Interface, Performance, and Demo), D2-1 (Vault Interface to Archival Storage)
- September 8: E3-1/2/3 (OS/Legion Security), F2-2 (AppLeS Demo)
- September 15: Discussion of each demo scheduled for Sept 24-26
- September 22: Last minute preparations for Sept 24-26 demos
- September 29: Demo de-briefing

Demonstrations:

- Demos and Presentations were made at the DARPA TIE facility from September 24-26
- A demonstration for the PTO Systems Development and Maintenance (SDM) contractors was given on July 10, 1997.

- A demonstration for the Executive Board of the American Intellectual Property Law Association (AIPLA) was given on July 10, 1997.
- A presentation was made to the Nuclear Regulatory Commission (NRC) on September 21, 1997.

4. Technical Results

The following is a list of tasks completed in FY97.Q4. Each task (or group of tasks) is followed by a summary of the findings and conclusions.

A6-1 - VRML Submission Analysis
A6-2 - Demonstrate VRML Content Building Software
http://www.sdsc.edu/DOCT/Publications/a6/a6.html

The A6 task group has been changed to a series of white papers that will evaluate the requirements of 3D content submission and make recommendations about supporting 3D content in the future. This white paper gives an introduction to 3D shape grammars and summarizes 3 major formats: DXF, RIB and VRML. Also included is an analysis of system constraints resulting from accepting 3D submissions.

A10-3 - Demonstrate Environmental Data in Archival Object Framework

The Interactive Repository of Environmental Data (IRED) is a web-based interface to metadata for environmental data repositories. IRED allows a user to seamlessly explore and download data from remote, geographically distributed data archives into a vis5D viewer using the SRB technology. A demo was given at the DOCT quarterly meeting in September, 1997.

A14 - Document classification using self-organizing maps

Research is being conducted on the generation of feature vectors for a particular US Patent classification category (364.550) using different vector sizes. Once a suitable set is defined, the feature vectors will be used to initiate a classification across the database.

B5-3 - Implementation of initial security model

Each of the DOCT partners has implemented the recommended items from the Policies, Standards and Guidelines (PSG) Security report in Task B5-4. This report is available at the DOCT URL.

B7-1 - USPTO provision of patent data

Tapes covering Jan 1975 to Jun 1997 have been received by SDSC and loaded into HPSS. In addition, a searchable database has been built which holds all of the patent metadata. Tapes covering Jul 1997 to Sep 1997 are being created by the USPTO and will be loaded into HPSS when received. As time allows, years 1971 to 1974 will be created by the USPTO and loaded into HPSS.

C5-1 - White paper: MDAS API
http://www.sdsc.edu/MDAS/SRBhello

The MDAS and DOCT projects have produced the Storage Resource Broker (SRB) that provides a uniform access mechanism to diverse and distributed data sources. The SRBhello suite contains

30 or more programs which demonstrate the SRB client library and an argument parsing package for command-line program development. These programs can be executed by a script that runs a battery of validation tests on an SRB installation. A release note is available under the MDAS URL.

C5-2 - MDAS Ticket API

The SDSC Encryption/Authentication (SEA) system was developed to provide authentication and encryption capabilities between two running processes communicating via TCP/IP sockets. In DOCT, the SEA is being used for communications between the SRB clients and servers.

C5-3 - Modify Legion to MDAS

Aspects of this task are dealt with in Tasks D1-4 and D2-3. Legion has been incorporated into MDAS, except that the metadata in the MDAS catalog is not used yet. The major benefit of this merger is that data can be migrated between different data resources.

C5-5 - MDAS Metadata for scheduling

Metadata is modeled and represented in the MDAS catalog to make it available for use in scheduling resources.

C5-6 - MDAS platform ports

The Storage Resource Broker software has been ported to the following platforms: AIX, Solaris, SunOS, SGI and Alpha. Storage systems handled include the UNIX file system, archival storage systems such as UNITREE and HPSS, and database Large Objects managed by various DBMS's including DB2 and Illustra.

D1-2 - Implement Vault Interface to Database
D2-2 - Implement Vault Interface to Archive

http://www.sdsc.edu/DOCT/Publications/d1-2/d1-2.html

Legion and the Storage Resource Broker (SRB) have been integrated by creating an SRB-based vault object (LegionVaultSRB). Since the SRB can access both archival storage systems and databases, Legion Vaults now also have that ability.

D1-4 - Demo of Data Access to Database
D2-3 - Demo of Data Access to Archive

An MPEG player was modified to use Legion to access files from databases and archival storage systems. The result was a movie shown in Virginia from an MPEG file in HPSS at SDSC. Legion can also stream data, so the movie could start playing before all of it was received.

F1-2 - Demo of queuing systems

Capabilities of the Cray Network Queuing Environment (NQE) were shown, including NQE job submission and monitoring and NQE data collection. This system provides risk mitigation software to supplement the Legion environment.

G3-1 - Update requirements and design based on Task G2-3 results
G3-2 - Update agent framework and develop new agents

http://www.sdsc.edu/DOCT/Publications/g3-1/g3-1.html

The agent framework continued to evolve to support the end-to-end patent submission process that was demonstrated Sep 24-26. We have also produced another view of the framework showing the workflow from the viewpoint of the various users. This is shown in Figure 4-1.

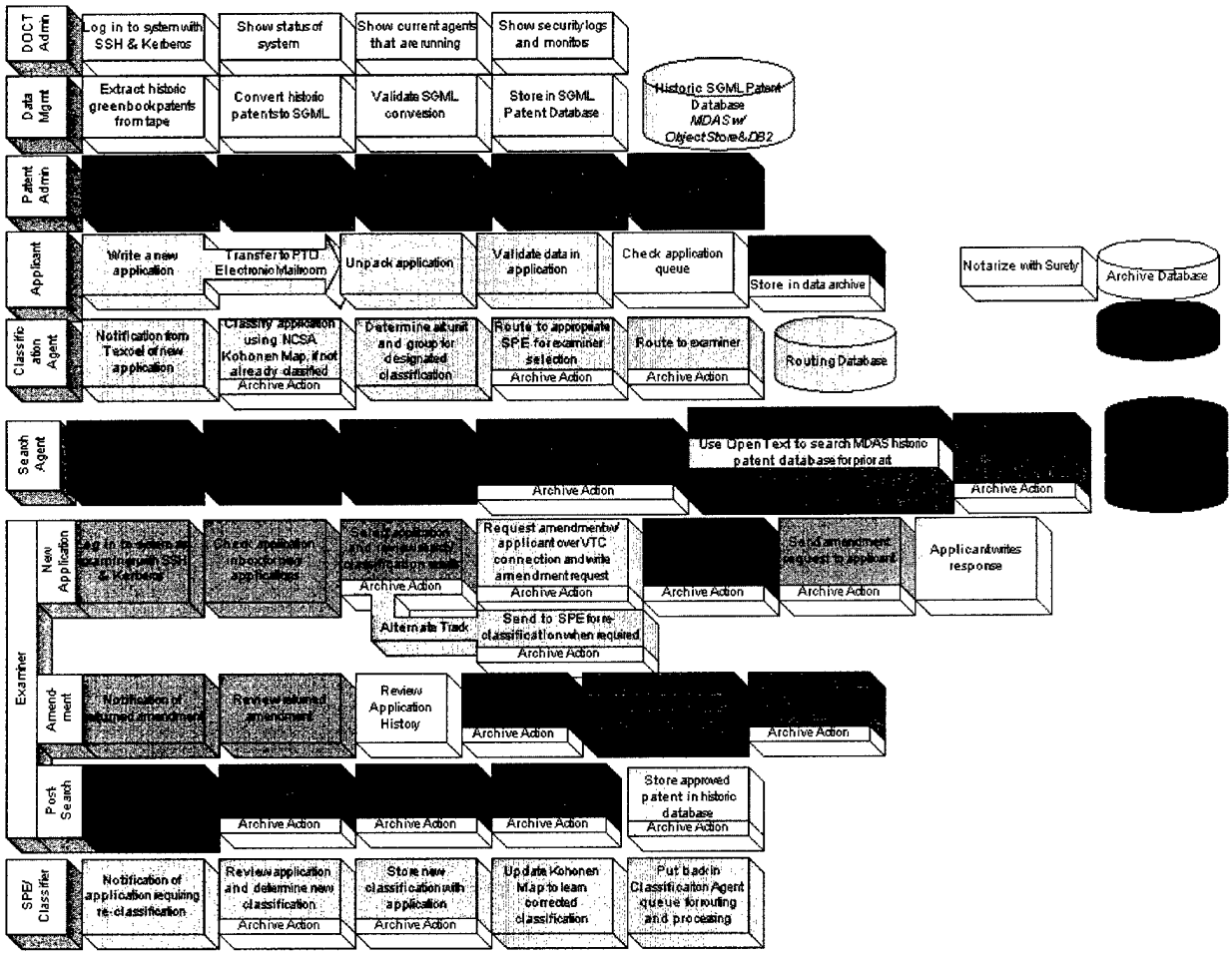

Figure 4-1: DOCT Architecture from User Viewpoints

H3-6 - Provide and coordinate VTC capability

We have installed and tested MBONE software at SDSC, SAIC and NCSA. The conclusion reached was that these tools were insufficient to get true multi-site VTC and the cost of a dedicated VTC setup at each site outweighed the benefits. For DOCT, we have continued to use weekly phone teleconferences for communication between the partners.

H3-7 - Provide and implement management software tools

The project management for deliverables has remained at the task level. The lead person for each task is responsible for managing the resources to produce a deliverable. When the task is done, the

deliverable(s) are forwarded to the project administrator to publish on the DOCT web page. For financial management, the monthly invoices from each partner, along with internal reports at SDSC, are summarized in an Excel spreadsheet made available at the DOCT URL.

H4-5 - DG5 Initial integrated system demonstrations
H4-6 - DG6 Workflow agent demonstrations

A set of demonstrations were performed on September 24-26 for the USPTO, DARPA and other agencies. The specific topics shown were:

- Authoring
- Surety Digital Notary
- Electronic Filing
- Validation
- Document Management with Agents
- Patent Data Loading into SGML format
- OpenText Search
- BRS Dataware Text Search
- Data Mining against the full database
- Security (SRB and SEA)
- NQE queuing system
- Interactive Repository of Environmental Data

4.1 Testbed implementation status

The DOCT testbed consists of hardware resources located at 8 sites coupled by software and networks spanning the country (see Figure 3-1). The testbed hardware implementation is complete, with some final optimizations to become available as listed:

SDSC - Raid upgrade to improve reliability of data import
Caltech, UVa, ODU - vBNS network connection

The software system implementation is substantially completed, with the installation of the following systems:

All sites - Legion 0.5 upgrade
NCSA, Caltech, SAIC - NQE installation

The Legion software will be upgraded to version 1.0 in November.

5. Special Comments

We are maintaining a Research and Development Plan & Schedule (RDPS) document for this project. The RDPS is a detailed, comprehensive overview of all DOCT tasks, including all deliverables (reports and demonstrations). This document is the authoritative source for our project goals, tasks, deliverables, schedule, and expected work loads. The RDPS is available via the World Wide Web at the URL:

http://www.sdsc.edu/DOCT/Publications/rdps-doc.html

We have also written a Concept of Operations (ConOps) document which describes the DOCT testbed in terms of its features and capabilities, its relevance to HPCC in the coming decade, and opportunities for other federal agencies and DARPA projects to participate and/or collaborate. This document specifies both ongoing research efforts and future R&D efforts relevant to multiple Federal Agency missions. The ConOps is available via the World Wide Web at the URL:

http://www.sdsc.edu/DOCT/Publications/conops/conops.html

Also note that final versions of the completed reports, progress reports and related documents can be found via the World Wide Web at:

http://www.sdsc.edu/DOCT/Publications.html